STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVII**, Number 4, 2012

PHONEMES VERSUS GEOMETRIC PROPERTIES IN CLUSTERING OF POEMS

MIHAIELA LUPEA AND DOINA TĂTAR

Abstract. This paper discusses the comparison between two kinds of features in clustering of some literary poems by the same author, the Romanian poet, Mihai Eminescu. Using *Precision*, *Recall*, *Rand Index*, *Relative Precision* and *Purity* measures we conclude that the topics of poems are better characterized by the phonemes as features than by geometric properties (described by six indicators: $V/N; A; \Lambda; Var(\Lambda), Gini; Var(Gini)$) of the rank-frequency sequence of word forms.

1. Introduction

This paper discusses the comparison between two kinds of features in clustering of some literary poems by the same author, the Romanian poet Mihai Eminescu. The *Gold Standard* (GS) of evaluation is a manual one, which divides the set of the longest 45 Eminescu's poems (see Appendix A for correspondence numbers - titles) into five big clusters, topic (content) focused:

- Love-general stories (tales): $\{10; 48; 51; 62; 64; 87; 93; 104; 109; 110; 117; 120; 129\}$;
- Love-personal stories: $\{8; 9; 13; 21; 38; 57; 68; 69; 94; 100; 143\}$;
- Philosophy-tales wisdom: $\{25; 34; 52; 58; 61; 70; 90; 106; 126; 127; 130\}$;
- Nature: $\{28; 91\}$;
- History-patriotism: $\{6; 47; 49; 50; 74; 95; 123; 128\}$.

For clustering the poems, these are represented using the space vector method and the Algorithm of Agglomerative Hierarchical Clustering ([2, 4]). The complete-link similarity between two clusters and the cosine similarity measure between two vectors are applied.

Received by the editors: 20.11.2012.

2010 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Text analysis*.

Key words and phrases. clustering, Rand Index, purity, phonemes.

	same cluster *Meth*	different cluster *Meth*
same cluster *GS*	A	C
different cluster *GS*	B	D

$P_{Meth} = \frac{A}{A+B}$ $R_{Meth} = \frac{A}{A+C}$	$A + B = \sum_{i=1}^{K} C^2_{\lvert w_i \rvert}$ $A + C = \sum_{i=1}^{K} C^2_{\lvert w_i^{GS} \rvert}$	$D = C^2_{\lvert M \rvert} - (A + B + C)$

Table 1. *Precision* and *Recall* measures

For the first clustering we considered as features the phonemes, building phonemes vectors corresponding to the phonetic transcription of the poems. Such a vector for a poem has 31 components containing the relative frequencies of the Romanian phonemes in that poem, as it is presented in Section 3.

The features used in the second clustering are geometric properties of the rank-frequency sequence of word forms in poems, expressed by vectors containing six indicators: V/N; A; Λ ; $\mathrm{Var}(\Lambda)$; *Gini*; Var(*Gini*)), introduced in [1] and described in Section 3.

The first method of evaluation of the clusterings is by establishing classical *Precision* and *Recall* measures, as reported to the Gold Standard (GS) clustering. The second method is *Rand Index* ([2]) and a *Relative Precision* as inspired from *Rand Index* algorithm (Section 2.2). The third method is the calculus of *Purity* ([2, 3]), Section 2.3.

In all these cases (excepting *Rand Index*) the conclusion is that the best indicators are the phonemes. The reason for these results seems to be the fact the most indicators (introduced in [1]) are based on words, and the words consists of phonemes. So, the phonemes unify and refine the words function. However, in this paper we worked with only a part of the indicators introduced in [1].

2. Evaluation of clustering

2.1. **Precision and Recall.** Let M be a set of elements. Two clustering methods are applied to M obtaining the same number K of clusters:

- an arbitrary method *Meth*, providing the clusters: $w_1, \ldots, w_K$;
- a manual method, providing the gold standard *GS* clustering: $w_1^{GS}, \ldots, w_K^{GS}$.

For a comparison of these two clusterings Table 1 is built. In the table we use C_T^2 to denote the binomial coefficient indexed by T and 2.

	same cluster *Meth*	different cluster *Meth*
same cluster *GS*	A	C
different cluster *GS*	B	D
$RI = \frac{A+D}{A+B+C+D}$	$A + B = \sum_{i=1}^{K} C^2_{\lvert w_i \rvert}$ $A + C = \sum_{j=1}^{J} C^2_{\lvert C_j \rvert}$	$D = C^2_{\lvert M \rvert} - (A + B + C)$

TABLE 2. *Rand Index* measure

The value of A represents the number of pairs of elements from M with the property: if a pair belongs to the same cluster obtained with *Meth*, it belongs also to the same cluster of *GS*.

The significance and the values of *B, C, D* are defined in an analogous way, deductible from the positions in Table 1. $A+B$ represents the number of pairs of elements situated in the same cluster obtained using *Meth*, and $A+C$ represents the number of pairs of elements situated in the same cluster of *GS*.

Precision: $P_{Meth} = \frac{A}{A+B}$ counts how many of the determined cases by *Meth* are correct.

Recall: $R_{Meth} = \frac{A}{A+C}$ counts how many of the correct cases are determined by *Meth*.

2.2. **Rand Index.** The set M is partitioned by some objective observations in J classes: $C_1, \ldots, C_J$. An arbitrary clustering method *Meth* is applied to M obtaining K clusters: $w_1, \ldots, w_K$.

The measure **Rand Index** (*RI*) penalizes both the False positive pairs (B) and the False negative pairs (C) according to Table 2.

Using *Rand Index* measure, we could obtain a method of a direct comparison of two clusterings *R1* and *R2* with the same number of clusters.

Rand Index of clustering R1 relative to R2, denoted by $RI_{R1,R2}$ expresses how good the clustering *R2* is, when a cluster (of *R2*) is considered a class: a cluster is calculated by a more or less good method, a class is judged by some objective reasons, thus a partition in classes is more exact than a partition in clusters. A similar significance has $RI_{R2,R1}$, expressing the quality of clustering *R1*.

$RI_{R1,R2} \leq RI_{R2,R1}$ means a better quality of the clustering *R1* than of the clustering *R2* (with *RI* measure), when the same similarity measures of clustering are used in *R1* and *R2*.

The method could be applied also for the case of *Precision*, namely, *Relative Precisions*: $P_{R1,R2}$ and $P_{R2,R1}$ could be calculated. $P_{R1,R2} \leq P_{R2,R1}$ means a better quality of the clustering *R1* than that of the clustering *R2*.

2.3. **Purity.** Let us suppose that we have K clusters: $w_1, \ldots, w_K$ and J classes: $C_1, \ldots, C_J$ for a set M of elements. The purity of a cluster w_k is calculated as:

$$Purity(w_k) = max_j\{n_{kj}\}/|w_k|$$

where $n_{kj} = |w_k \cap C_j|$.

The index $j_k^* = \text{argmax}_j\ n_{kj}$ determines the majority class of the cluster w_k denoted by $C_{j_k^*}$.

The $Purity(w_k)$ is the number of elements provided by the majority class of the cluster w_k over the cardinal of the cluster. The higher the contribution of the majority class, the higher the purity of a cluster.

The *Purity* of a clustering is the weighted sum of the purities of all clusters:

$$Purity = \sum_{k=1}^{K} Purity(w_k) \times \text{weight}(w_k)$$

where $weight(w_k) = |w_k|/|M|$.

3. A case study - clustering of Eminescu's poems

In this section we apply the theory from the previous section using as M the set of the 45 longest poems of Eminescu (Appendix A). The poems are represented using the vector space method, where the vectors are:

(1) numeric vectors of 31 components containing the relative frequencies of the Romanian phonemes in the poem, describing the content of the poem in a phonetic manner. The phonemes correspond to the vowels (in number of 7), consonants (in number of 18) and 6 groups of letters('ce', 'ci', 'ge', 'gi', 'ch', 'gh'). The letter 'x' is decomposed in two phonemes [c]+[s].

For example, the statistics for the poem *Memento mori* (90) are:

- total phonemes number: 46433;
- vowels number: 21494;
- consonants number: 24939 (including the groups of letters);
- the vector of occurrences for all 31 phonemes(in this order: vowels, consonants, the groups of letters) is: (3995, 5035, 4593, 1852, 3024, 1767, 1228, 512, 1614, 1827, 501, 385, 39, 70, 2491, 1431, 3370, 1341, 4072, 1831, 681, 2433, 362, 563, 402, 428, 262, 95, 103, 117, 9).

For the phoneme 'a', with 3995 occurences in the poem, its relative frequency in the category of vowels is computed as: 3995/21494 = **0.1859**.

The relative frequency in the category of consonants for the phoneme 'b', with 512 occurences in the poem is computed as: 512/24939 = **0.0205**.

The vector of phonemes for *Memento mori* is: (**0.1859**, 0.2343, 0.2137, 0.0862, 0.1407, 0.0822, 0.0571, **0.0205**, 0.0647, 0.0733, 0.0201, 0.0154, 0.0016, 0.0028, 0.0999, 0.0574, 0.1351, 0.0538, 0.1633, 0.0734, 0.0273, 0.0976, 0.0145, 0.0226, 0.0161, 0.0172, 0.0105, 0.0038, 0.0041, 0.0047, 4.0E-4).

(2) numeric vectors of six components corresponding to some indicators: V/N; A; Λ ; Var(Λ); *Gini*; Var(*Gini*), which describe geometric properties of the rank-frequency sequence of word forms in poems ([1]).

The significance of the indicators is the following: V is the vocabulary size (words) of the text, N is the text length (the total number of words in the text), A (adjusted modulus) is an index of vocabulary richness.

As regarding Λ indicator, this is introduced as a normalization of L, the length of the arc beginning at $f(1)$ and ending at $f(V)$, $\Lambda = L/N * (Log(N)))$. *Gini*'s coefficient is connected with the cumulative relative frequencies which form an arc running from (0,0) and touching the bisector in (1,1). The magnitude of the area between the bisector and this arc yields *Gini*'s coefficient. The expressions for Var(Λ) and Var(*Gini*) are also introduced first time in ([1]).

For example, the vector for the poem *Memento mori* (90) is: (0.365906068, 0.9311, 1.6175, 0.000068, 0.5717, 0.000033).

To obtain five clusters (like in *Gold Standard*) of Eminescu's poems we used the Algorithm of Agglomerative Hierarchical Clustering ([2, 4]) (or bottom-up clustering algorithm), the complete-link similarity between two clusters and the cosine similarity measure between two vectors.

In the bottom-up clustering algorithm we begin with a separate cluster for each poem and we continue by grouping the most similar clusters until we obtain a specific number of clusters (here five clusters).

For the cosine similarity measure between the vectors $V_1 = (a_1, a_2, ..., a_n)$ and $V_2 = (b_1, b_2, ..., b_n)$ the well known formula is used:

$$sim(V_1, V_2) = cos(V_1, V_2) = \frac{\sum_{i=1}^{i=n} a_i * b_i}{\sqrt{\sum_{i=1}^{i=n} a_i^2} \times \sqrt{\sum_{i=1}^{i=n} b_i^2}}$$

The complete-link similarity between two clusters $C1$ and $C2$ represents the similarity of two least similar members of the two clusters:

Precision	*Recall*	*Rand Index*	*Relative Precision*
$P_{R1} = 0.2997$	$R_{R1} = 0.5622$	$RI_{R1} = 0.6161$	$P_{R1,R2} = 0.2088$
$P_{R2} = 0.2392$	$R_{R2} = 0.2811$	$RI_{R2} = 0.6383$	$P_{R2,R1} = 0.3231$

TABLE 3. Measures for *R1* and *R2* clusterings

$$sim(C1, C2) = min\{sim(V_i, V_j)|V_i \in C1 \text{ } and \text{ } V_j \in C2\}.$$

The clustering $R1$ corresponds to the representation (1):

- w_1^{R1}:{8(2); 9(2); 21(2); 38(2); 57(2); 68(2); 69(2); 70(3); 94(2); 123(5)};
- w_2^{R1}:{6(5); 10(1); 13(2); 25(3); 34(3); 47(5); 48(1); 49(5); 50(5); 51(1); 52(3); 61(1); 62(1); 64(1); 74(5); 87(1); 90(3); 95(5); 109(1); 110(1); 117(1);126(3); 127(3); 128(5); 129(3); 130(2); 143(2)};
- w_3^{R1}:{28(4); 91(4); 93(1); 104(1); 120(1)};
- w_4^{R1}:{58(3); 100(2)};
- w_5^{R1}:{106(3)}.

The clustering $R2$ corresponds to the representation (2):

- w_1^{R2}: {52(3); 90(3)};
- w_2^{R2}: {6(5); 10(1); 87(1); 95(5); 109(1); 128(5); 130(3)};
- w_3^{R2}: {34(3); 58(3); 61(3); 91(4); 126(3); 129(1)};
- w_4^{R2}: {9(2); 13(2); 21(2); 25(3); 28(4); 48(1); 49(5); 50(5); 62(1); 64(1); 69(2); 93(1); 94(2); 104(1); 106(3); 110(1); 117(1); 127(3); 143(2)};
- w_5^{R2}:{8(2); 38(2); 47(5); 51(1); 57(2); 68(2); 70(3); 74(5); 100(2); 120(1); 123(5)}.

The numbers in brackets represent the manual assignation for the poems of one of the five clusters corresponding to *Gold Standard* clustering (see Introduction).

$R1$ and $R2$ are compared applying the measures of *Precision, Recall, Rand Index, Relative Precision*, and *Purity* and the results are reported in Table 3.

For computing the purities of *R1* and *R2* we consider that *GS* clustering represents the set of predefined classes. Table 4 contains the values of purities for all clusters of *R1* and $R2$, and also the overall purities for these clusterings.

The overall *Purity* for $R1$, $Purity_{R1} = 0.8 \times 0.22 + 0.37 \times 0.6 + 0.6 \times 0.11 + 0.5 \times 0.04 + 1 \times 0.02 = 0.504$.

clusters of *R1*	w_1^{R1}	w_2^{R1}	w_3^{R1}	w_4^{R1}	w_5^{R1}	
Purity	0.8	0.37	0.6	0.5	1	$Purity_{R1} = 0.504$

clusters of *R2*	w_1^{R2}	w_2^{R2}	w_3^{R2}	w_4^{R2}	w_5^{R2}	
Purity	1	0.42	0.66	0.36	0.45	$Purity_{R2} = 0.438$

TABLE 4. *Purity* measure for *R1* and *R2* clusterings

The overall *Purity* for $R2$, $Purity_{R2} = 1 \times 0.04 + 0.42 \times 0.15 + 0.66 \times 0.13 + 0.36 \times 0.42 + 0.45 \times 0.24 = 0.438$.

From Table 3 and Table 4 we can conclude:

(1) Both *Precision* and *Recall* are better in the case of *R1* clustering than in the case of *R2* clustering: $P_{R1} \geq P_{R2}$ and $R_{R1} \geq R_{R2}$.
(2) According to *Rand Index* measure the results are better for *R2* than for *R1*: $RI_{R1} \leq RI_{R2}$.
(3) In a direct comparison of *R1* and *R2* clusterings using *Relative Precision*, the quality of *R1* is better than that of *R2*: $P_{R1,R2} \leq P_{R2,R1}$.
(4) As $Purity_{R1} \geq Purity_{R2}$, we can say again that the clustering *R1* is of a better quality than $R2$.

4. Conclusions

In this paper *Precision, Recall, Rand Index, Relative Precision* and *Purity* evaluation measures are used to compare the impact of different features of poems in the topic-focused clustering of the 45 longest Eminescu's poems. Excepting *Rand Index*, all the other measures suggest that the phonemes as features characterize better the topic (content) of the poems than geometric properties (described by the indicators: $V/N; A; \Lambda; Var(\Lambda), Gini; Var(Gini)$) of the rank-frequency sequence of word forms.

References

[1] Popescu, I.I., Čech, R., Altmann, G.: "The Lambda-structure of Texts"'. Studies in quantitative linguistics 10, RAM-Verlag, 2011.

[2] Manning, C., Raghavan, P., Schutze, H.: "Introduction to Information Retrieval", Cambridge University Press, 2008.

[3] Mihalcea, R., Radev, D.: "Graph-based Natural language Processing and Infromation Retrieval", Cambridge University Press, 2011.

[4] Tatar, D., Serban, G.: "Word clustering in QA systems", Studia Universitatis Babes-Bolyai, Seria Informatica 2003, 1, pp. 23–33.

Appendix A. The correspondence between numbers and titles of poems

No.	*Poem*	*No.*	*Poem*
(6)	Andrei Mureşanu	(8)	Aveam o muză
(9)	Basmul ce i l-aş...	(10)	Călin
(13)	Când crivăţul cu iarna...	(21)	Copii eram noi amândoi
(25)	Cugetările sărmanului...	(28)	Dacă treci râul Selenei
(34)	Demonism	(38)	Despărţire
(47)	Dumnezeu şi om	(48)	Eco
(49)	Egipetul	(50)	Epigonii
(51)	Făt-Frumos din tei	(52)	Feciorul de impărat fără...
(57)	Ghazel	(58)	Glossa
(61)	Impărat şi proletar	(62)	In căutarea Şeherezadei
(64)	Inger şi demon	(68)	Iubită dulce, o, mă lasă
(69)	Iubitei	(70)	Junii corupti
(74)	La moartea lui Heliade	(87)	Luceafărul
(90)	Memento mori	(91)	Miradoniz
(93)	Mitologicale	(94)	Mortua est!
(95)	Mureşanu	(100)	Nu mă-nţelegi
(104)	O călărire în zori	(106)	O, adevăr sublime...
(109)	Odin şi poetul	(110)	Ondina (Fantazie)
(117)	Povestea teiului	(120)	Pustnicul
(123)	Rugăciunea unui dac	(126)	Scrisoarea I
(127)	Scrisoarea II	(128)	Scrisoarea III
(129)	Scrisoarea IV	(130)	Scrisoarea V
(143)	Venere şi Madonă		

Babeş-Bolyai University, Department of Computer Science, M. Kogălniceanu St., 400084 Cluj-Napoca, Romania

E-mail address: lupea,dtatar@cs.ubbcluj.ro